Appendix

Duy Truong

10/25/2018

library(qwraps2)  
library(GGally)

## Loading required package: ggplot2

library("Hmisc")

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(ggcorrplot)  
library(ggplot2)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:qwraps2':  
##   
## logit

library(readxl)  
library(tidyverse)

## ── Attaching packages ──────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ tibble 1.4.2 ✔ purrr 0.2.5  
## ✔ tidyr 0.8.2 ✔ dplyr 0.7.7  
## ✔ readr 1.3.0 ✔ stringr 1.3.1  
## ✔ tibble 1.4.2 ✔ forcats 0.3.0

## ── Conflicts ─────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::recode() masks car::recode()  
## ✖ purrr::some() masks car::some()  
## ✖ dplyr::src() masks Hmisc::src()  
## ✖ dplyr::summarize() masks Hmisc::summarize()

Housing\_data <- read\_excel("Housing data.xlsx")

Data transformation

Housing\_data$taxModified<-as.numeric(Housing\_data$TAX)

## Warning: NAs introduced by coercion

NA Analysis

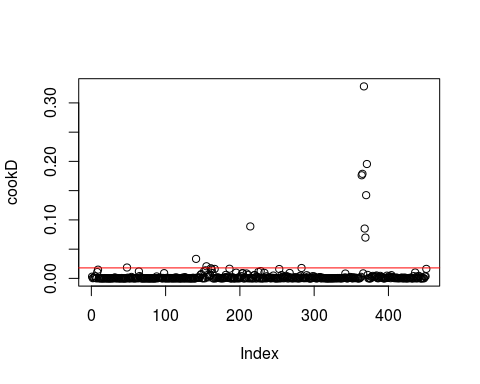
#create a new dataframe with omitted na data  
HousingUpdate <- na.omit(Housing\_data)

Identify outlier

#fit a model without data modification and na data  
fit<- lm(MEDV ~ CRIM + ZN + CHAS + RM + AGE + DIS + log(RAD) + taxModified + PTRATIO + LSTAT, data=HousingUpdate)  
summary(fit)

##   
## Call:  
## lm(formula = MEDV ~ CRIM + ZN + CHAS + RM + AGE + DIS + log(RAD) +   
## taxModified + PTRATIO + LSTAT, data = HousingUpdate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0503 -2.5625 -0.5434 1.7280 29.6378   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.218536 4.234644 3.122 0.001918 \*\*   
## CRIM -0.198071 0.178376 -1.110 0.267429   
## ZN 0.043422 0.013809 3.144 0.001777 \*\*   
## CHAS 2.517552 0.842427 2.988 0.002961 \*\*   
## RM 5.324072 0.454678 11.710 < 2e-16 \*\*\*  
## AGE -0.014902 0.012900 -1.155 0.248643   
## DIS -1.062846 0.182879 -5.812 1.19e-08 \*\*\*  
## log(RAD) 1.755045 0.519172 3.380 0.000788 \*\*\*  
## taxModified -0.010445 0.002994 -3.488 0.000535 \*\*\*  
## PTRATIO -0.605265 0.120736 -5.013 7.78e-07 \*\*\*  
## LSTAT -0.562998 0.057724 -9.753 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.665 on 440 degrees of freedom  
## Multiple R-squared: 0.7258, Adjusted R-squared: 0.7196   
## F-statistic: 116.5 on 10 and 440 DF, p-value: < 2.2e-16

#apply cook's distance to identify outliers in the model  
cookD <-cooks.distance(fit)  
#plot the cook distance to measure how many outliers  
plot(cookD)  
abline(h = 4\*mean(cookD), col="red")



#introduce new data with outliers  
outlier <- as.numeric(names(cookD)[(cookD > 4\*mean(cookD))])

New model with removed outliers

#new dataframe without outlier and na data  
HousingClean <-HousingUpdate[-outlier, ]  
#count how many rows are left after cutting the outliers to get the observation numbers  
nrow(HousingClean)

## [1] 440

#Final model  
fitUpdate<- lm(MEDV ~ CRIM + CHAS+ZN + RM + AGE + DIS + log(RAD) + taxModified + PTRATIO + log(LSTAT), data=HousingClean)

Correlation table

HousingCleanCorr <- HousingClean[,c(13,1,2,3,4,6,7,8,9,11,14)]  
cor(HousingCleanCorr)

## LSTAT CRIM ZN INDUS CHAS  
## LSTAT 1.00000000 0.49364409 -0.42814838 0.6083506 0.025250255  
## CRIM 0.49364409 1.00000000 -0.27604953 0.5643438 -0.011648130  
## ZN -0.42814838 -0.27604953 1.00000000 -0.5117165 -0.044522781  
## INDUS 0.60835064 0.56434381 -0.51171645 1.0000000 0.065012204  
## CHAS 0.02525025 -0.01164813 -0.04452278 0.0650122 1.000000000  
## RM -0.63324723 -0.14569058 0.31539709 -0.3777320 0.043014453  
## AGE 0.61839196 0.43803760 -0.55776285 0.6035989 0.100975407  
## DIS -0.46837259 -0.44202146 0.65791122 -0.6589992 -0.103975567  
## RAD 0.38589693 0.89796773 -0.26205253 0.5036892 -0.003933441  
## PTRATIO 0.33100183 0.30927607 -0.36318710 0.3089013 -0.123203100  
## MEDV -0.72826731 -0.39569542 0.36758914 -0.4803185 0.109269821  
## RM AGE DIS RAD PTRATIO  
## LSTAT -0.63324723 0.6183920 -0.4683726 0.385896926 0.3310018  
## CRIM -0.14569058 0.4380376 -0.4420215 0.897967729 0.3092761  
## ZN 0.31539709 -0.5577629 0.6579112 -0.262052528 -0.3631871  
## INDUS -0.37773197 0.6035989 -0.6589992 0.503689157 0.3089013  
## CHAS 0.04301445 0.1009754 -0.1039756 -0.003933441 -0.1232031  
## RM 1.00000000 -0.1916309 0.1369244 -0.101015649 -0.3472475  
## AGE -0.19163086 1.0000000 -0.7262312 0.346626632 0.1874684  
## DIS 0.13692439 -0.7262312 1.0000000 -0.366329838 -0.1352325  
## RAD -0.10101565 0.3466266 -0.3663298 1.000000000 0.3757768  
## PTRATIO -0.34724753 0.1874684 -0.1352325 0.375776848 1.0000000  
## MEDV 0.83404568 -0.3542431 0.1980811 -0.324609650 -0.5311498  
## MEDV  
## LSTAT -0.7282673  
## CRIM -0.3956954  
## ZN 0.3675891  
## INDUS -0.4803185  
## CHAS 0.1092698  
## RM 0.8340457  
## AGE -0.3542431  
## DIS 0.1980811  
## RAD -0.3246097  
## PTRATIO -0.5311498  
## MEDV 1.0000000

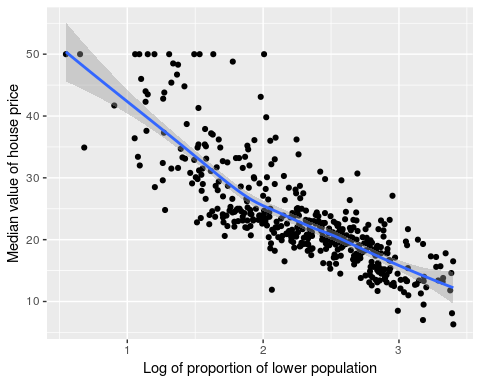
Condition check

1. Linear model

Correlation of each variable to response variable

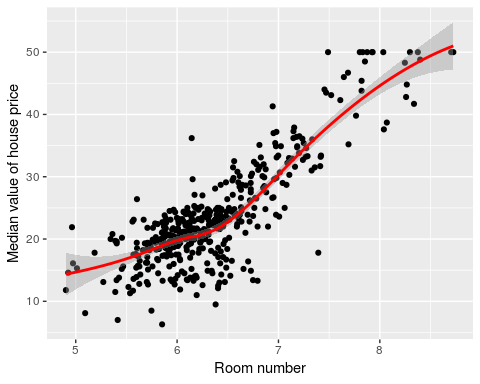
#Graph showing relationship between median house price and the proportion population  
ggplot(data=HousingClean) +  
 geom\_point(mapping = aes(x = log(LSTAT) , y = MEDV))+  
 geom\_smooth(mapping = aes(x = log(LSTAT) , y = MEDV))+  
 labs(x='Log of proportion of lower population', y='Median value of house price')

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



#Graph showing relationship between median house price and the room number  
ggplot(data=HousingClean) +  
 geom\_point(mapping = aes(x =RM, y = MEDV))+  
 geom\_smooth(mapping = aes(x = RM , y = MEDV), col='red')+  
 labs(x='Room number',y='Median value of house price')

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



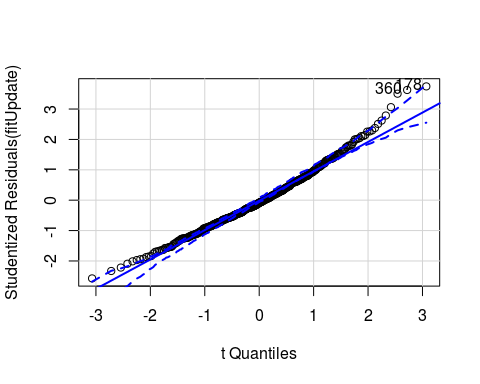
1. Multicollearity check

vif(fitUpdate)

## CRIM CHAS ZN RM AGE DIS   
## 3.943572 1.047634 2.365694 2.647519 3.018934 3.017190   
## log(RAD) taxModified PTRATIO log(LSTAT)   
## 3.339355 4.082016 1.459986 4.094605

All values are under an acceptable range 3. Multivariate normality

qqPlot(fitUpdate)

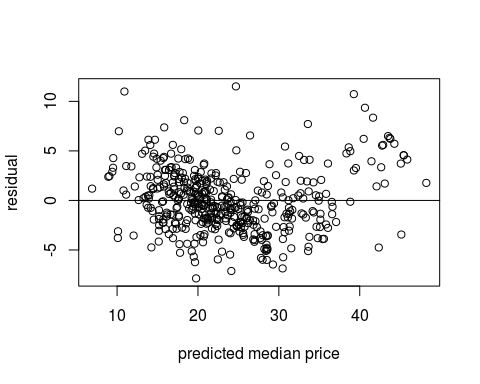


## [1] 178 360

The residuals follows a normal distribution, which satisfies the multivariate condition

4.Homoscedasticity

plot(fitUpdate$fit, fitUpdate$residual, ylab="residual", xlab="predicted median price")  
abline(0,0)



The graph shows no pattern in the distribution, therefore satisfying the homoscedasticity condition.